(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
12 April 2001 (12.04.2001)

PCT

(10) International Publication Number
WO 01/26092 A2

(51) International Patent Classification[7]: G10L 15/18

(21) International Application Number: PCT/IB00/01539

(22) International Filing Date: 6 October 2000 (06.10.2000)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/157,874 6 October 1999 (06.10.1999) US

(71) Applicant: LERNOUT & HAUSPIE SPEECH PRODUCTS N.V. [BE/BE]; Flanders Language Valley 50, B-8900 Ieper (BE).

(72) Inventors: FINKE, Michael; 1172 Murray Hill Avenue, Pittsburgh, PA 15217 (US). FRITSCH, Juergen; Lachnerstr, 23, 76131 Karlsruhe (DE). KOLL, Detleff; 6608 Dalzell PL, #2, Pittsburgh, PA 15217 (US). WAIBEL, Alex; 619 Windsor Avenue, Pittsburgh, PA 15221 (US).

(74) Agent: FROUD, Clive; Elkington and Fife, Prospect House, 8 Pembroke Road, Sevenoaks, Kent TN13 1XR (GB).

(81) Designated States *(national)*: AU, CA, JP.

(84) Designated States *(regional)*: European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published:
— *Without international search report and to be republished upon receipt of that report.*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: ATTRIBUTE-BASED WORD MODELING

(57) Abstract: An attribute-based speech recognition system is described. A speech preprocessor receives input speech and produces a sequence of acoustic observations representative of the input speech. A database of context-dependent acoustic models characterize a probability of a given sequence of sounds producing the sequence of acoustic observations. Each acoustic model includes phonetic attributes and suprasegmental non-phonetic attributes. A finite state language model characterizes a probability of a given sequence of words being spoken. A one-pass decoder compares the sequence of acoustic observations to the acoustic models and the language model, and outputs at least one word sequence representative of the input speech.

WO 01/26092 A2

Attribute-Based Word Modeling

Field of the Invention

The invention relates to automatic speech recognition, and more particularly, to word models used in a speech recognition system.

Background Art

Automatic speech recognition (ASR) systems do not effectively address variations in word pronunciation. Typically, ASR dictionaries contain few alternative pronunciations for each entry. In natural speech, however, words rarely follow their citation forms. This failure to capture an important source of variability can cause recognition errors, particularly in normal conversational speech.

The automatic inference of pronunciation variation has been explored using phonetically transcribed corpora. Unfortunately, increasing the number of dictionary entry variants based on a pronunciation model also increases the confusability between dictionary entries, and thus often leads to an actual performance decrease.

Speaking mode has been considered to reduce confusability by probabilistically weighting alternative pronunciations depending on the speaking style. See F. Alleva, X. Huang, M.-Y. Hwang, *Improvements on the Pronunciation Prefix Tree Search Organization*, Proc. Int. Conf. on Acoustics, Speech and Signal Processing, Atlanta, GA, pp. 133 - 136, May 1996 (incorporated herein by reference). This approach uses pronunciation modeling and acoustic modeling based on a wide range of observables such as speaking rate; duration; and syllabic, syntactic, and semantic structure—contributing factors that are subsumed in the notion of speaking mode. See, *e.g.*, M. Ostendorf, B. Byrne, M. Bacchiani, M. Finke, A. Gunawardana, K. Ross, S. Roweis, E. Shriberg, D. Talkin, A.Waibel, B. Wheatley, and T. Zeppenfeld, *Systematic Variations in Pronunciation via a Language-Dependent Hidden Speaking Mode*, in International Conference on

Spoken Language Processing, Philadelphia, USA, 1996 (incorporated herein by reference).

Just as the phonetic representation of careful speech is a schematization of articulatory and acoustic events, a phonetic transcription of relaxed informal speech by its nature is a simplification. Pronunciation models implementing purely phonological mappings generate phonetic transcriptions that underspecify durational and spectral properties of speech. Reduced variants as predicted by a pronunciation model ought to be phonetically homophonous—*e.g.*, the fast variant of "support" being pronounced as /s/p/o/r/t/ is phonetically homophonous with "sport"). But for to create such homophony, not only should the unstressed vowels be deleted, but the durations of the remaining phones also should take the same values as in words not derived from fast speech vowel reduction. Similarly, fast speech intervocalic voicing in a word like "faces" cannot be precisely represented as /f/ey/z/ih/z/—phonetically homophonous with "phases"—unless both the voice value of the fricative as well as the durational relationship between the stressed vowel and the fricative have changed.

Brief Description of the Drawings

The present invention will be more readily understood by reference to the following detailed description taken with the accompanying drawings, in which:

Figure 1 illustrates a prefix search tree according to a representative embodiment of the present invention showing roots, nodes, leaves, and single phone word nodes (stubs).

Figure 2 illustrates the heap structure of a root node.

Figure 3 illustrates the heap structure of a leaf node.

Figure 4 illustrates the heap structure of a stub.

Detailed Description of Specific Embodiments

The foregoing suggests that capturing the complex variability of conversational speech with purely phone-based speech recognizers is virtually impossible. Embodiments of the present invention generalize phonetic speech transcription to an attribute-based representation that integrates supra-segmental non-phonetic features. A pronunciation model is trained to augment an attribute transcription by marking possible pronunciation effects, which are then taken into account by an acoustic model induction algorithm. A finite state machine single-prefix-tree, one-pass, time-synchronous decoder is used to decode highly spontaneous speech within this new representational framework.

In representative embodiments, the notion of context is broadened from a purely phonetic concept to one based on a set of speech attributes. The set of attributes incorporates various features and predictors such as dialect, gender, articulatory features (e.g. vowel, high, nasal, shifted, stress, reduced), word or syllable position (e.g. word begin/end, syllable boundary), word class (e.g pause, function word), duration, speaking rate, fundamental frequencies, HMM state (e.g. begin/middle/end state), etc. This approach affects all levels of modeling within the recognition engine, from the way words are represented in the dictionary, through pronunciation modeling and duration modeling, to acoustic modeling. This leads to strategies to efficiently decode conversational speech within the mode dependent modeling framework.

A word is transcribed as a sequence of instances ($\iota_0\iota_1 \ldots \iota_k$) which are bundles of instantiated attributes (i.e. attribute-value pairs). Each attribute can be either binary, discrete (i.e. multi-valued), or continuous valued. For example, the filled pause "um" is transcribed by a single instance ι consisting of truth values for the following binary attributes (pause, nasal, voiced, labial ...).

The instance-based representation allows for a more detailed modeling of pronunciation effects as observed in sloppy informal speech. Instead of predicting

an expected phonetic surface form based on a purely phonetic context, the canonical instance-based transcription is probabilistically augmented. A pronunciation model predicts instances for the set of attributes. Instead of mapping from one phone sequence to another, the pronunciation model is trained to predict pronunciation effects:

$$p(\iota_k' \mid ..\iota_{k-1}[\iota_k]\iota_{k+1}...)$$

Pronunciation variants are derived by augmenting the initial transcription by the predicted instances:

$$\iota_0\iota_1..\iota_k \mapsto (\iota_0 \oplus \iota_0')(\iota_1 \oplus \iota_1')...(\iota_k \oplus \iota_k')$$

which are then weighted by a probability:

$$p(\iota_0'\iota_1'..\iota_k') = \frac{1}{Z}\prod_{\kappa=0}^{k} p(\iota_\kappa' \mid ..\iota_{\kappa-1}[\iota_\kappa]\iota_{\kappa+1}...)$$

where Z is a normalizing constant.

Predicting pronunciation variation, as described above, by augmenting the phonetic transcription by expected pronunciation effects avoids potentially homophonous representation of variants (see, e.g., M. Finke and A. Waibel, *Speaking Mode Dependent Pronunciation Modeling in Large Vocabulary Conversational Speech Recognition*, in Proceedings of Eurospeech-97, September 1997, incorporated. herein by reference).

The original transcription is preserved, and the duration and acoustic model building process exploit the augmented annotation. Decision trees are grown to induce a set of context dependent duration and acoustic models. The induction algorithm allows for questions with respect to all attributes defined in the transcription. Thus, starting from the augmented transcription, context dependent modeling means that the acoustic models derived depend on the phonetic context, pronunciation effects, and speaking mode-related attributes. This leads to a much tighter coupling of pronunciation modeling and acoustic modeling because model induction takes the pronunciation predictors into

account as well as acoustic evidence.

For coherence of training, testing, and rescoring results, a corresponding LVCSR decoder should handle finite state grammar decoding, forced alignment of training transcripts, large vocabulary statistical grammar decoding, and lattice rescoring. One typical embodiment uses a single-prefix-tree time-synchronous one-pass decoder that represents the underlying recognition grammar by an abstract finite state machine. To have reasonable efficiency in a one-pass decoder, the dictionary is represented by a pronunciation prefix tree as described in H. Ney, R. Haeb-Umbach, B.-H. Tran, M. Oerder, "*Improvement In Beam Search For 10000-word Continuous Speech Recognition*," IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 9-12, 1992, incorporated herein by reference.

Two problems can result from this representation. First, if the tree is reentrant, then only the single best history may be considered at word transitions at each time t. Second, the application of the grammar score is delayed since the identity of the word is only known at the leaves of the tree. To deal with the first problem, a priority heap may represent alternative linguistic theories in each node of the prefix tree as described in previously cited Alleva, Huang, and Hwang. The heap maintains all contexts whose probabilities are within a certain threshold, thus avoiding following only the single best local history. The threshold and the heap policy have the benefit of allowing different more or less aggressive search techniques by effectively controlling hypothesis merging. In contrast to the tree copying process employed by other recognizers, the heap approach is more dynamic and scalable.

The language model is implemented in the decoder as an abstract finite state machine. The exact nature of the underlying grammar remains transparent to the recognizer. The only means to interact with a respective language model is through the following set of functions in which FSM is a finite state machine-

based language model:

FSM.initial()—Returns the initial state of the FSM.

FSM.arcs(state)—Returns all arcs departing from a given state. An arc consists of the input label (recognized word), the output label, the cost, and the next state. Finite state machines are allowed to be non-deterministic, i.e. there can be multiple arcs with the same input label.

FSM.cost(state)—Returns the exit cost for a given state to signal whether or not a state is a final state.

This abstraction of the language model interface makes merging of linguistic theories a straightforward and well-defined task to the decoder: two theories fall into the same congruence class of histories and can be merged if the state indices match. The finite state machine is designed to return which theories can be merged. One advantage of this division of labor is that the decoder can decode grammars of any order without any additional implementation effort.

To deal with filler words, i.e. words that are not modeled by a particular FSM grammar (these are typically pauses such as silence and noises), the decoder may virtually add a self loop with a given cost term to each grammar state. As a result any number of filler words can be accepted/recognized at each state of the finite state machine.

One typical embodiment provides a set of different instantiations of the finite state machine interfaces that are used in different contexts of training, testing or rescoring a recognizer:

- **Finite State Grammar Decoding**—Of course, the FSM interface may explicitly define a finite state grammar. Besides its use in command-and-control, this application can be used in training the recognizer. In M. Finke and A. Waibel, *Flexible Transcription Alignment*, in ASRU, pages 34–40, Santa Barbara, CA, December 1997, we showed that, when dealing with unreliable transcripts of training data, a significant gain in word accuracy can be achieved by training

from probabilistic transcription graphs instead of the raw transcripts. Typical embodiments allow for decoding of right recursive rule grammars by simulating an underlying heap to deal with recursion. The transcription graphs of the Flexible Transcription Alignment (FTA) paradigm may be expressed in the decoder by a probabilistic rule grammar. Thus, forced alignment of training data is basically done through decoding these utterance grammars.

- **N-gram Decoding**—Statistical n-gram language models are not explicitly represented as a finite state machine. Instead, a finite state machine wrapper is built around n-gram models. The state index codes the history such that FSM.arcs(state) can retrieve all the language model scores required from the underlying n-gram tables. This implies that the FSM is not minimized and the state space is the vocabulary to the power of the order of the n-gram model.
- **Lattice Rescoring**—Lattices are finite state machines, too. So, rescoring a word graph using a different set of acoustic models and a different language model is feasible by decoding along lattices and on-the-fly composition of finite state machines.

Grammar probabilities should be incorporated into the search process as early as possible so that tighter pruning thresholds can be used for decoding. Within the finite state machine abstraction, lookahead techniques can be generalized to any kind of FSM based language model. See, *e.g.* S. Ortmanns, A. Eiden, H. Ney, and N. Coenen, *Look-Ahead Techniques for Fast Beam Search*, in Proceedings of the ICASSP'97, pages 1783--1786, Munich (Germany), 1997, incorporated herein by reference. For each state, the decoder needs to derive—on demand—a cost tree that reports for each node of the prefix tree what the best language model score is going to be for all words with a given prefix. For a trigram-based FSM, the lookahead tree will thus be a trigram lookahead; for fourgrams, a fourgram lookahead; and for finite state grammars, the lookahead will be a projection of all words allowed at a given grammar state. In order to

compute finite state machine lookahead trees efficiently on demand, several techniques can be combined:

- Lookahead trees may be saved in an aging cache as they are computed to avoid recomputing the tree in subsequent frames.
- The size of the cache and the number of steps to compute the tree can be reduced by precomputing a new data structure from the prefix tree: the cost tree. The cost tree represents the cost structure in a condensed form, and turns the rather expensive recursive procedure of finding the best score in the tree into an iterative algorithm.
- Each heap element, hypothesis, or tree copy has the current FSM lookahead score attached. When a hypothesis is expanded to the next node and the corresponding lookahead tree has been removed from the cache, the tree will not be recomputed. Instead, the lookahead probability of the prefix is propagated forward ("lazy cache" evaluation).

Typical embodiments use polyphonic within-word acoustic models, but triphone acoustic models across word boundaries. To incorporate crossword modeling in a single-prefix-tree decoder, the context dependent root and leaf nodes are dealt with. Instead of having context dependent copies of the prefix tree, each root node may be represented as a set of models, one for each possible phonetic context. The hypotheses of these models are merged at the transition to the within word units (fan-in). As a compact means of representing the fan-in of root nodes and the corresponding fan-out of leaf nodes, the notion of a multiplexer was developed. A multiplexer is a dual map that maps instances ι to the index of a unique hidden Markov model for the context of ι:

$$\mathrm{mpx}(\iota): \iota \mapsto i \in \{0,1,\ldots,N_{mpx}\}$$

$$\mathrm{mpx}[i]: i \mapsto m \in \{m_0, m_1, \ldots, m_{N_{mpx}}\}$$

where $m_0, m_1, \ldots, m_{N_{mpx}}$ are unique models. The set of multiplexer models can be

precomputed based on the acoustic modeling decision tree and the dictionary of the recognizer. Figure 1 shows the general organization of a multiplexer-based prefix search tree showing various type of nodes including a root node **10**, internal node **12**, leaf node **14**, and single phone word node **16** (also called a stub).

To model conversational speech, multiplexers are particularly useful since the augmented attribute representation of words leads to an explosion in the number of crossword contexts. Because multiplexers map to unique model indices, they basically implement a compression of the fan-in/out and a technique to address the context dependent model by the context instance ι.

The heap structure of a root node, **10** in Fig. 1, is shown in Figure 2. The root node **10** represents the first attribute instance of words in terms of a corresponding multiplexer **20**. The structure also includes for each node a finite state machine grammar state **22** and corresponding state index **24**. Cost structure **26** contains a finite state machine lookahead score for the node. Score structure **28** contains a total best score of hypotheses, the acoustic score plus expected FSM cost. Heap policy only merges hypotheses that have the same history or linguistic theory, and whose final instances ι_a and ι_b map to the same context dependent word initial model: $\mathrm{mpx}(\iota_a) = \mathrm{mpx}(\iota_b)$. This means that the heap is used to keep track of different contexts, the FSM state (representing the linguistic context), as well as acoustic contexts. In word internal nodes **12**, only hypotheses found to be in the same FSM state are collapsed.

For every word there is a leaf node, **14** in Fig. 1, the heap structure of which is illustrated by Figure 3. A multiplexer describes the leaf node fan-out, and each heap element represents the complete fan-out for a given grammar state.

Figure 4 illustrates the heap structure for a single-phone instance word node or stub, **16** in Fig. 1. Words consisting of only one phone are represented by a multiplexer of multiplexers. Depending on the left context of the word, this stub

multiplexer returns a multiplexer representing the right-context dependent fan-out of this word. The heap policy is the same as for root nodes, and each heap element represents the complete fan-out as for leaf nodes.

In addition to the acoustic and the word end beam for pruning the acoustics, two heap related controls may also be used: (1) the maximum number of heap elements can be bounded, and (2) there can be a beam to prune hypotheses within a heap against each other. The number of finite state machine states expanded at each time t can be constrained as well (topN threshold).

Acoustic model evaluation is sped up by means of gaussian selection through Bucket Box Intersection (BBI) and by Dynamic Frame Skipping (DFS). Thus, acoustic models are reevaluatedonly provided the acoustic vector changed significantly from time *t* to time *t+1*. A threshold on the Euclidean distance is defined to trigger reevaluation of the acoustics. To avoid skipping too many consecutive frames, only one skip at a time may be taken—after skipping one frame, the next one must be evaluated.

To assess the performance of a representative embodiment of the decoder under tight realtime constraints, an evaluation test started from a Switchboard recognizer trained on human-to-human telephone speech. The acoustic front end computed 42 dimensional feature vectors consisting of 13 mel-frequency cepstral coefficients plus log power and their first and second derivatives. Cepstral mean and variance normalization as well as vocal tract length normalization were used to compensate for channel and speaker variation. The recognizer consisted of 8000 pentaphonic Gaussian mixture models. A 15k word recognition vocabulary and approximately 30k dictionary variants generated by a mode dependent pronunciation model were used for decoding. Without MLLR adaptation, and decoded with a Switchboard trigram language model trained on 3.5 million words, the base performance at 100xRT was 37% word error rate (run-on, one-pass recognition on NIST Eval'96). Groups participating in recent NIST

evaluations reported decoding times in the order of 300 realtime factors (which included multiple adaptation passes).

Table 1 shows the first word accuracy results of our Switchboard recognizer at around ten times realtime. This shows tight pruning in the context of highly confusable Switchboard speech. TopN=10 means that only 10 finite state machine states were expanded per frame. DFS indicates Dynamic Frame Skipping, and BBI indicates Bucket Box Intersection.

Condition	RT	WER%
Baseline	100	37
Tight Beams, topN=10	12	43.8
Tight Beams, topN=10, DFS	7	45.6
Tight Beams, topN=10, DFS, BBI	5	49.8

Table 1

Embodiments of the invention may be implemented in any conventional computer programming language. For example, preferred embodiments may be implemented in a procedural programming language (*e.g.*, "C") or an object oriented programming language (*e.g.*, "C++"). Alternative embodiments of the invention may be implemented as pre-programmed hardware elements, other related components, or as a combination of hardware and software components.

Embodiments can be implemented as a computer program product for use with a computer system. Such implementation may include a series of computer instructions fixed either on a tangible medium, such as a computer readable medium (*e.g.*, a diskette, CD-ROM, ROM, or fixed disk) or transmittable to a computer system, via a modem or other interface device, such as a communications adapter connected to a network over a medium. The medium may be either a tangible medium (*e.g.*, optical or analog communications lines) or a medium implemented with wireless techniques (*e.g.*, microwave, infrared or other transmission techniques). The series of computer instructions embodies all

or part of the functionality previously described herein with respect to the system. Those skilled in the art should appreciate that such computer instructions can be written in a number of programming languages for use with many computer architectures or operating systems. Furthermore, such instructions may be stored in any memory device, such as semiconductor, magnetic, optical or other memory devices, and may be transmitted using any communications technology, such as optical, infrared, microwave, or other transmission technologies. It is expected that such a computer program product may be distributed as a removable medium with accompanying printed or electronic documentation (*e.g.*, shrink wrapped software), preloaded with a computer system (*e.g.*, on system ROM or fixed disk), or distributed from a server or electronic bulletin board over the network (*e.g.*, the Internet or World Wide Web). Of course, some embodiments of the invention may be implemented as a combination of both software (*e.g.*, a computer program product) and hardware. Still other embodiments of the invention are implemented as entirely hardware, or entirely software (*e.g.*, a computer program product).

Although various exemplary embodiments of the invention have been disclosed, it should be apparent to those skilled in the art that various changes and modifications can be made which will achieve some of the advantages of the invention without departing from the true scope of the invention.

What is claimed is:

1. An attribute-based speech recognition system comprising:
a speech pre-processor that receives input speech and produces a sequence of acoustic observations representative of the input speech;
a database of context-dependent acoustic models that characterize a probability of a given sequence of sounds producing the sequence of acoustic observations, each acoustic model including phonetic attributes and suprasegmental non-phonetic attributes;
a finite state language model that characterizes a probability of a given sequence of words being spoken; and
a one-pass decoder that compares the sequence of acoustic observations to the acoustic models and the language model, and outputs at least one word sequence representative of the input speech.

2. A speech recognition system according to claim 1, wherein the suprasegmental non-phonetic attributes include speaking rate.

3. A speech recognition system according to claim 1, wherein the suprasegmental non-phonetic attributes include phone durations.

4. A speech recognition system according to claim 1, wherein the suprasegmental non-phonetic attributes include dialect.

5. A speech recognition system according to claim 1, wherein the suprasegmental non-phonetic attributes include gender.

6. A speech recognition system according to claim 1, wherein the

suprasegmental non-phonetic attributes include fundamental frequencies.

7. A speech recognition system according to claim 1, wherein the suprasegmental non-phonetic attributes include hidden Markov model state.

8. A speech recognition system according to claim 1, wherein the suprasegmental non-phonetic attributes include word class.

9. A speech recognition system according to claim 1, wherein the suprasegmental non-phonetic attributes include articulatory features.

10. A speech recognition system according to claim 9, wherein the articulatory features include stress.

11. A speech recognition system according to claim 1, wherein the suprasegmental non-phonetic attributes possess discrete values.

12. A speech recognition system according to claim 11, wherein the discrete values are binary.

13. A speech recognition system according to claim 1, wherein the suprasegmental non-phonetic attributes possess continuous values.

14. A speech recognition system according to claim 1, wherein the suprasegmental non-phonetic attributes include syllabic structure.

15. A speech recognition system according to claim 1, wherein the suprasegmental non-phonetic attributes include syntactic structure.

16. A speech recognition system according to claim 1, wherein the suprasegmental non-phonetic attributes include semantic structure.

17. A speech recognition system according to claim 1, wherein the decoder further comprises:
a probabilistic pronunciation model that characterizes possible pronunciation effects,
wherein an acoustic model induction algorithm augments the acoustic models with the pronunciation model.

18. A speech recognition system according to claim 1, wherein the decoder is a single-prefix-tree decoder.

19. A speech recognition system according to claim 18, wherein the prefix tree includes nodes and leaves, and for each node and leaf a priority heap represents alternative linguistic theories.

20. A speech recognition system according to claim 19, wherein the heap maintains all contexts within a selected threshold probability.

21. A speech recognition system according to claim 19, wherein lookahead cost trees are used to determine for every node a best language model score for all words with a given prefix.

22. A speech recognition system according to claim 21, wherein lookahead trees are saved in an aging cache to avoid recomputing for subsequent frames.

23. A speech recognition system according to claim 21, wherein each heap element has a current lookahead score attached.

24. A speech recognition system according to claim 19, wherein the decoder further comprises:
a multiplexer to represent fan-in of root nodes and fan-out of leaf nodes.

25. A speech recognition system according to claim 24, wherein the prefix tree includes root nodes to represent a first attribute instance of words for a given multiplexer.

26. A speech recognition system according to claim 24, wherein the prefix tree includes word internal nodes wherein the decoder collapses alternative hypotheses only if the alternative hypotheses are in the same finite machine state.

27. A speech recognition system according to claim 24, wherein each word has a leaf node, and the multiplexer describes the fan-out such that each heap element represents a complete fan-out for a given grammar state.

28. A speech recognition system according to claim 24, wherein a single phone word is represented by a multiplexer of multiplexers that, depending on a left context, returns a multiplexer representing a right context-dependent fan-out of the single phone word.

29. A speech recognition system according to claim 1, wherein the decoder is time synchronous.

30. A speech recognition system according to claim 1, wherein the decoder uses decision trees to induce a set of context-dependent duration and acoustic models.

31. A speech recognition system according to claim 1, wherein the language model uses n-gram models in a finite state machine wrapper.

32. A speech recognition system according to claim 1, wherein the decoder uses finite state recognition lattices that enable rescoring a word graph with alternative word models or language models.

33. A speech recognition system according to claim 1, wherein the decoder uses a bucket box intersection technique.

34. A speech recognition system according to claim 1, wherein the decoder uses a dynamic frame skipping technique.

35. An attribute-based method of speech recognition comprising:
pre-processing input speech to produce a sequence of acoustic observations representative of the input speech;
characterizing, with context-dependent acoustic models, a probability of a given sequence of sounds producing the sequence of acoustic observations, each acoustic model including phonetic attributes and suprasegmental non-phonetic attributes;
characterizing, with a finite state language model, a probability of a given sequence of words being spoken; and

comparing, with a one-pass decoder, the sequence of acoustic observations to the acoustic models and the language model, and outputs at least one word sequence representative of the input speech.

36. A method according to claim 35, wherein the suprasegmental non-phonetic attributes include speaking rate.

37. A method according to claim 35, wherein the suprasegmental non-phonetic attributes include phone durations.

38. A method according to claim 35, wherein the suprasegmental non-phonetic attributes include dialect.

39. A method according to claim 35, wherein the suprasegmental non-phonetic attributes include gender.

40. A method according to claim 35, wherein the suprasegmental non-phonetic attributes include fundamental frequencies.

41. A method according to claim 35, wherein the suprasegmental non-phonetic attributes include hidden Markov model state.

42. A method according to claim 35, wherein the suprasegmental non-phonetic attributes include word class.

43. A method according to claim 35, wherein the suprasegmental non-phonetic attributes include articulatory features.

44. A method according to claim 9, wherein the articulatory features include

stress.

45. A method according to claim 35, wherein the suprasegmental non-phonetic attributes possess discrete values.

46. A method according to claim 11, wherein the discrete values are binary.

47. A method according to claim 35, wherein the suprasegmental non-phonetic attributes possess continuous values.

48. A method according to claim 35, wherein the suprasegmental non-phonetic attributes include syllabic structure.

49. A method according to claim 35, wherein the suprasegmental non-phonetic attributes include syntactic structure.

50. A method according to claim 35, wherein the suprasegmental non-phonetic attributes include semantic structure.

51. A method according to claim 35, wherein the comparing further comprises:
characterizing, with a probabilistic pronunciation model , possible pronunciation effects, and
augmenting the acoustic models with the pronunciation model using an acoustic model induction algorithm.

52. A method according to claim 35, wherein the comparing uses a single-prefix-tree decoder.

53. A method according to claim 52, wherein the prefix tree includes nodes

and leaves, and for each node and leaf a priority heap represents alternative linguistic theories.

54. A method according to claim 53, wherein the priority heap maintains all contexts within a selected threshold probability.

55. A speech recognition system according to claim 53, wherein lookahead cost trees are used to determine for every node a best language model score for all words with a given prefix.

56. A method according to claim 55, wherein lookahead cost trees are saved in an aging cache to avoid recomputing for subsequent frames.

57. A speech recognition system according to claim 55, wherein each heap element has a current lookahead score attached.

58. A method according to claim 53, wherein the comparing further includes representing, with a multiplexer, fan-in of root nodes and fan-out of leaf nodes.

59. A method according to claim 58, wherein the prefix tree includes root nodes to represent a first attribute instance of words for a given multiplexer.

60. A method according to claim 58, wherein the prefix tree includes word internal nodes in which the decoder collapses alternative hypotheses only if the alternative hypotheses are in the same finite machine state.

61. A method according to claim 58, wherein each word has a leaf node and

the multiplexer describes the fan-out such that each heap element represents a complete fan-out for a given grammar state.

62. A method according to claim 58, wherein a single phone word is represented by a multiplexer of multiplexers that, depending on a left context, returns a multiplexer representing a right context-dependent fan-out of the single phone word.

63. A method according to claim 35, wherein the comparing includes using a time synchronous decoder.

64. A method according to claim 35, wherein the comparing includes using decision trees to induce a set of context-dependent duration and acoustic models.

65. A method according to claim 35, wherein the language model uses n-gram models in a finite state machine wrapper.

66. A method according to claim 35, wherein the comparing includes using finite state recognition lattices that enable rescoring a word graph with alternative word models or language models.

67. A method according to claim 35, wherein the comparing includes using a bucket box intersection technique.

68. A method according to claim 35, wherein the comparing includes using a dynamic frame skipping technique.

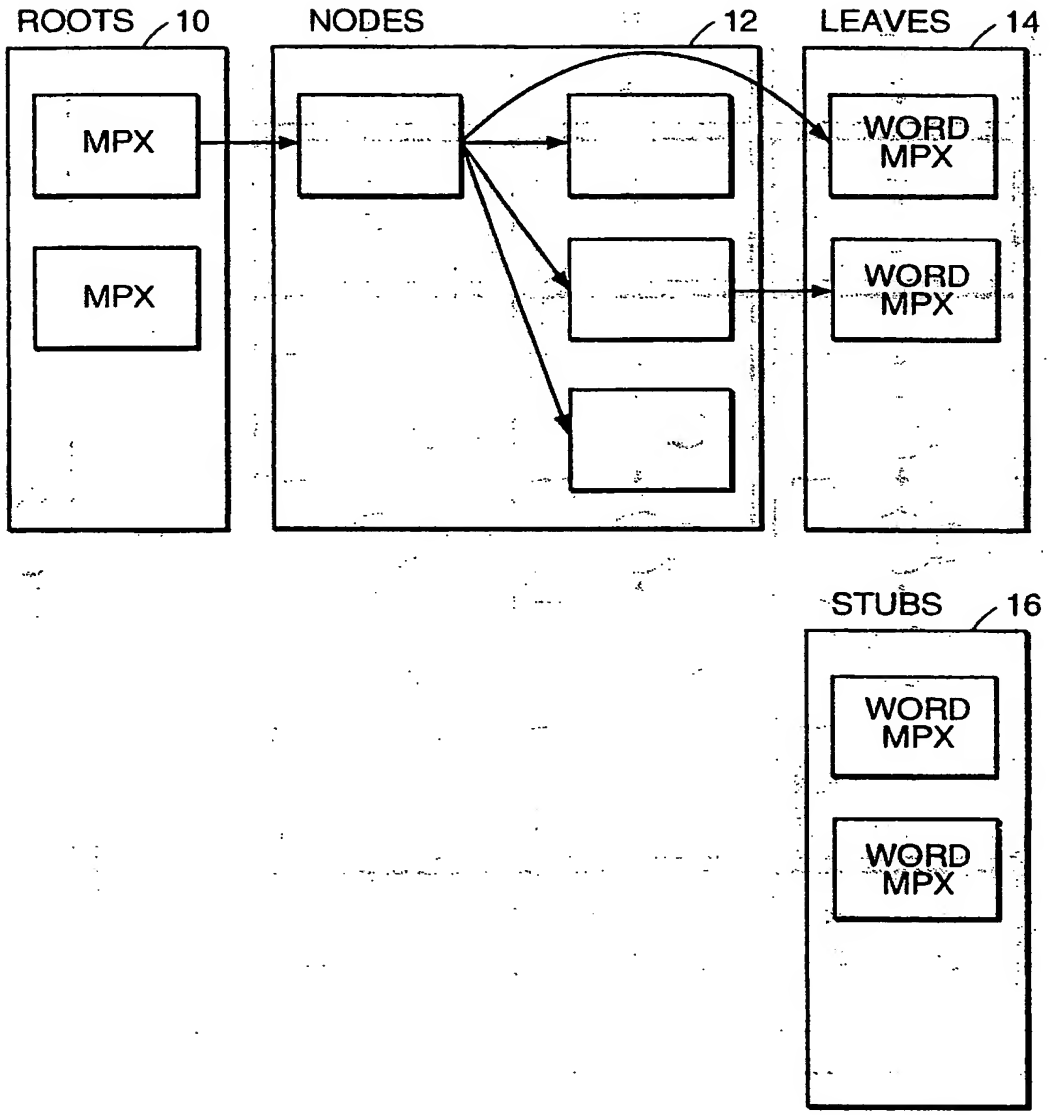

FIG. 1

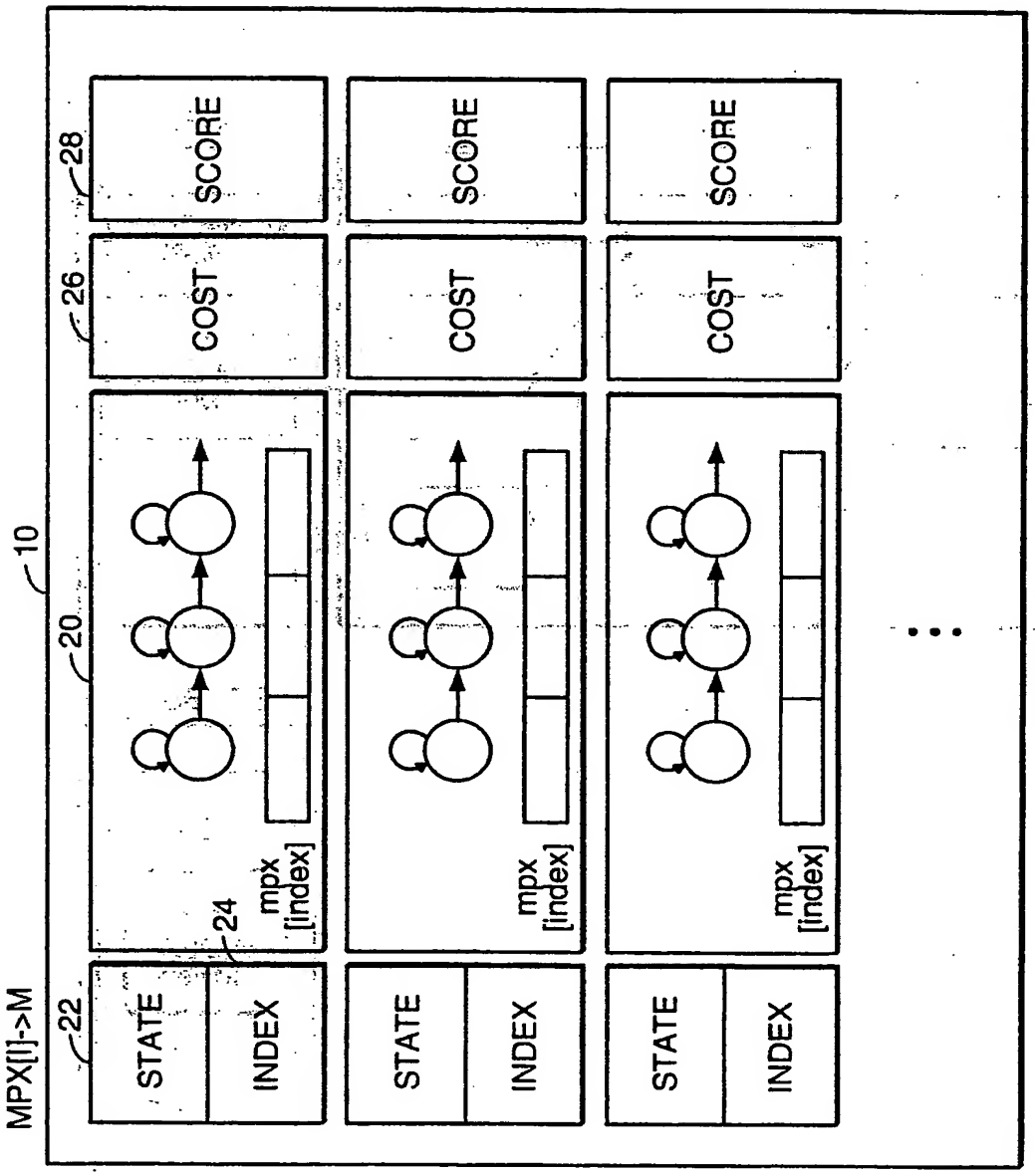

FIG. 2

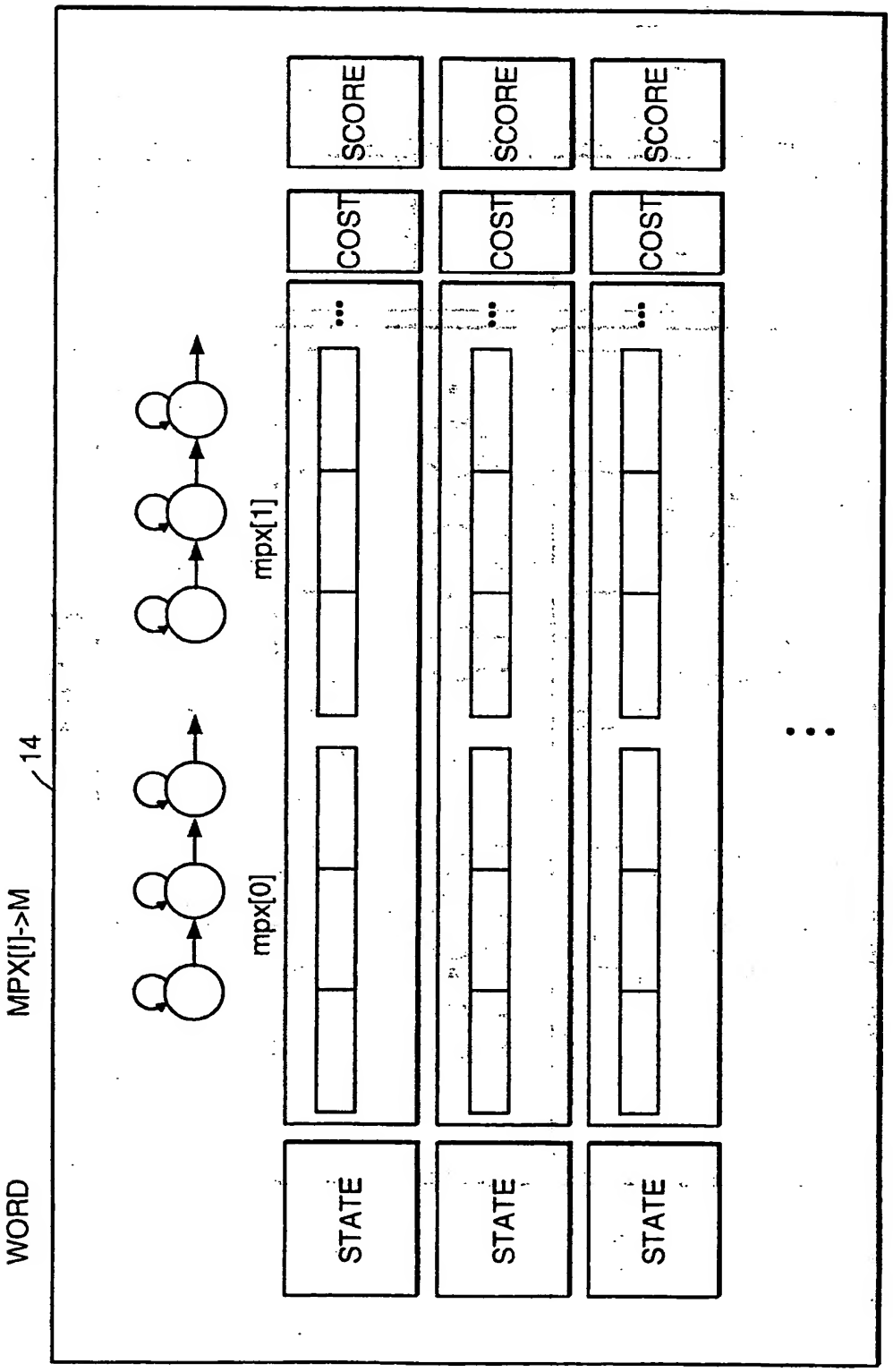

FIG. 3

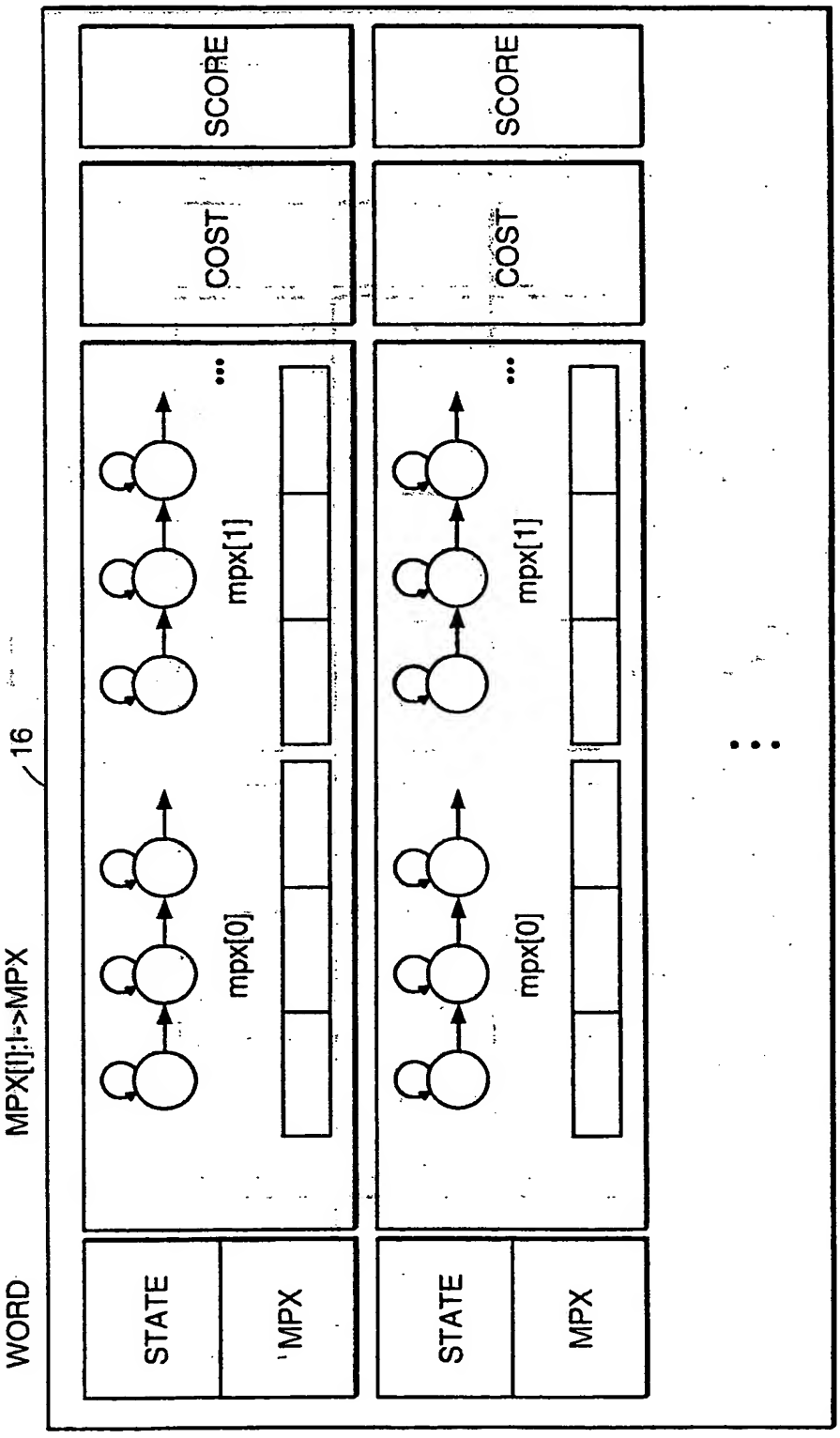

FIG. 4

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
12 April 2001 (12.04.2001)

PCT

(10) International Publication Number
WO 01/026092 A3

(51) International Patent Classification[7]: G10L 15/18, 15/08, 15/02

(21) International Application Number: PCT/IB00/01539

(22) International Filing Date: 6 October 2000 (06.10.2000)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/157,874 6 October 1999 (06.10.1999) US

(71) Applicant: LERNOUT & HAUSPIE SPEECH PRODUCTS N.V. [BE/BE]; Flanders Language Valley 50, B-8900 Ieper (BE).

(72) Inventors: FINKE, Michael; 1172 Murray Hill Avenue, Pittsburgh, PA 15217 (US). FRITSCH, Juergen; Lachnerstr, 23, 76131 Karlsruhe (DE). KOLL, Detleff; 6608 Dalzell PL, #2, Pittsburgh, PA 15217 (US). WAIBEL, Alex; 619 Windsor Avenue, Pittsburgh, PA 15221 (US).

(74) Agent: FROUD, Clive; Elkington and Fife, Prospect House, 8 Pembroke Road, Sevenoaks, Kent TN13 1XR (GB).

(81) Designated States *(national)*: AU, CA, JP.

(84) Designated States *(regional)*: European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published:
— *with international search report*

(88) Date of publication of the international search report:
22 May 2003

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: ATTRIBUTE-BASED WORD MODELING

(57) Abstract: An attribute-based speech recognition system is described. A speech preprocessor receives input speech and produces a sequence of acoustic observations representative of the input speech. A database of context-dependent acoustic models characterize a probability of a given sequence of sounds producing the sequence of acoustic observations. Each acoustic model includes phonetic attributes and suprasegmental non-phonetic attributes. A finite state language model characterizes a probability of a given sequence of words being spoken. A one-pass decoder compares the sequence of acoustic observations to the acoustic models and the language model, and outputs at least one word sequence representative of the input speech.

WO 01/026092 A3

INTERNATIONAL SEARCH REPORT

Internati pplication No
PCT/IB 00/01539

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G10L15/18 G10L15/08 G10L15/02

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
INSPEC, EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	FINKE ET AL.: "Modeling and efficient decoding of large vocabulary conversational speech" EUROSPEECH'99, vol. 1, 5 - 9 September 1999, pages 467-470, XP002168070 Budapest, Hungary the whole document ---	1,18-28, 32-35, 52-62, 66-68
Y	RENALS S ET AL: "Start-synchronous search for large vocabulary continuous speech recognition" IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, SEPT. 1999, IEEE, USA, vol. 7, no. 5, pages 542-553, XP002159651 ISSN: 1063-6676 abstract --- -/--	2-17, 29-31, 36-51, 63-65

[X] Further documents are listed in the continuation of box C. [ ] Patent family members are listed in annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier document but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
22 May 2001	29/06/2001

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Quélavoine, R

Form PCT/ISA/210 (second sheet) (July 1992)

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	SUAUDEAU N ET AL: "An efficient combination of acoustic and supra-segmental informations in a speech recognition system" ICASSP-94. 1994 IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING (CAT. NO.94CH3387-8), PROCEEDINGS, ADELAIDE, SA, AUSTRALIA, 19-22 APRIL 1, pages I/65-8 vol.1, XP002159652 1994, New York, NY, USA, IEEE, USA ISBN: 0-7803-1775-0 abstract	2,3,7, 36,37,41
Y	WAGNER M: "Speaker characteristics in speech and speaker recognition" TENCON '97 BRISBANE - AUSTRALIA. PROCEEDINGS OF IEEE TENCON '97. IEEE REGION 10 ANNUAL CONFERENCE. SPEECH AND IMAGE TECHNOLOGIES FOR COMPUTING AND TELECOMMUNICATIONS (CAT. NO.97CH36162), TENCON '97 BRISBANE - AUSTRALIA, page 626 vol.2 XP002159653 1997, New York, NY, USA, IEEE, USA ISBN: 0-7803-4365-4 abstract	4,5,38, 39
Y	WANG H -M ET AL: "COMPLETE RECOGNITION OF CONTINUOUS MANDARIN SPEECH FOR CHINESE LANGUAGE WITH VERY LARGE VOCABULARY BUT LIMITED TRAINING DATA" PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING (ICASSP),US,NEW YORK, IEEE, 9 May 1995 (1995-05-09), pages 61-64, XP000657931 ISBN: 0-7803-2432-3 abstract	6,8,40, 42
Y	ERLER K ET AL: "HMM REPRESENTATION OF QUANTIZED ARTICULATORY FEATURES FOR RECOGNITION OF HIGHLY CONFUSIBLE WORDS" PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING (ICASSP),US,NEW YORK, IEEE, vol. CONF. 17, 23 March 1992 (1992-03-23), pages 545-548, XP000341204 ISBN: 0-7803-0532-9 abstract	9,11-13, 43,45-47

-/--

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	ANASTASAKOS A ET AL: "Duration modeling in large vocabulary speech recognition" INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING,XX,XX, vol. 1, 9 May 1995 (1995-05-09), pages 628-631, XP002123829 abstract	10,44
Y	DATABASE INSPEC 'Online! INSTITUTE OF ELECTRICAL ENGINEERS, STEVENAGE, GB; DELMONTE R: "Linguistic tools for speech recognition and understanding" Database accession no. 4199465 XP002159657 abstract & SPEECH RECOGNITION AND UNDERSTANDING. RECENT ADVANCES, TRENDS AND APPLICATIONS. PROCEEDINGS OF THE NATO ADVANCED STUDY INSTITUTE, CETRARO, ITALY, 1-13 JULY 1990, pages 481-485, 1992, Berlin, Germany, Springer-Verlag, Germany ISBN: 3-540-54032-6	14,48
Y	LLORENS D ET AL: "ACOUSTIC AND SYNTACTICAL MODELING IN THE ATROS SYSTEM" PHOENIX, AZ, MARCH 15 - 19, 1999,NEW YORK, NY: IEEE,US, 15 March 1999 (1999-03-15), pages 641-644, XP000900202 ISBN: 0-7803-5042-1 abstract	15,49
Y	MERGEL D ET AL: "CONSTRUCTION OF LANGUAGE MODELS FOR SPOKEN DATABASE QUERIES" INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH & SIGNAL PROCESSING. ICASSP,US,NEW YORK, IEEE, vol. CONF. 12, 1 April 1987 (1987-04-01), pages 844-847, XP000758092 abstract	16,50
Y	BYRNE W ET AL: "PRONUNCIATION MODELLING USING A HAND-LABELLED CORPUS FOR CONVERSATIONAL SPEECH RECOGNITION" IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING,US,NEW YORK, NY: IEEE, vol. CONF. 23, 12 May 1998 (1998-05-12), pages 313-316, XP000854578 ISBN: 0-7803-4429-4 abstract	17,51

-/--

INTERNATIONAL SEARCH REPORT

Internatl Application No
PCT/IB 00/01539

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	NEY H ET AL: "Dynamic programming search for continuous speech recognition" IEEE SIGNAL PROCESSING MAGAZINE, SEPT. 1999, IEEE, USA, vol. 16, no. 5, pages 64-83, XP002159654 ISSN: 1053-5888 the whole document	29-31, 63-65
A	MYOUNG-WAN KOO ET AL: "A new decoder based on a generalized confidence score" INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, XX, XX, vol. 1, 1998, pages 213-216, XP002123828 abstract	1,3,35, 37

Form PCT/ISA/210 (continuation of second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/IB 00/01539

Box I Observations where certain claims were found unsearchable (Continuation of item 1 of first sheet)

This International Search Report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the International Application that do not comply with the prescribed requirements to such an extent that no meaningful International Search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box II Observations where unity of invention is lacking (Continuation of item 2 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

see additional sheet

1. ☒ As all required additional search fees were timely paid by the applicant, this International Search Report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this International Search Report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this International Search Report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest.

☒ No protest accompanied the payment of additional search fees.

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

This International Searching Authority found multiple (groups of) inventions in this international application, as follows:

1. Claims: 1-17,29-51,63-68

use of supra-segmental attributes

1.1. Claims: 29,63
a time synchronous decoder

1.2. Claims: 30,64
use of a decision tree for context dependent duration and acoustic models

1.3. Claims: 32,66
use of finite state recognition lattices for rescoring a word graph

1.4. Claims: 33,67
use of bucket box intersection

1.5. Claims: 34,68
use of dynamic frame skipping

1.6. Claims: 11-13,45-47
binary, discrete or continuous values

2. Claims: 18-28,52-62

use of a single prefix tree decoder

Please note that all inventions mentioned under item 1, although not necessarily linked by a common inventive concept, could be searched without effort justifying an additional fee.